G·32·6BN
2

IMPROVING ACCURACY OF GENE PREDICTION PROGRAMS OF THE GENEMARK FAMILY BY MEANS OF GENOME SEGMENTATION

RO3 Fogarty TW005899-01A1 program.

PS Project #: 32066BN

Final report

Development of the Computational & Network Infrastructure for the research partners in Russia

Network installations & accessibility: Direct high speed internet connection was established between the Moscow and the Atlanta groups of investigators. To increase the speed of internet connection at the EIMB Moscow site the Network Infrared Bridge was purchased to connect EIMB with the network hub nearby. A remote access to powerful computational facilities at Georgia Tech was arranged for the Moscow group. This connection allowed to access to databases in Atlanta and Moscow. Programs of the GeneMark family were installed at the Moscow site.

Travel: Dr. Borodovsky (Georgia Tech) visited EIMB in July 2003, December 2004 and October 2006. Dr. Makeev (EIMB) visited Georgia Tech in May 2003 and February 2004. Dr. Roytberg (EIMB) visited Georgia Tech in October 2005.

Fund transfer to the partners in Russia: In total, the financial support for the Moscow group in amount of $54,000 was distributed through the years as follows

$19,800 in 2003
$7,200 in 2004
$27,000 in 2006

Equipment for Moscow partners
The total of $25,840 was spent for equipment, of which
$19,800 in 2003; $1,200 in 2004; $4,840 in 2006

The following equipment was purchased by Moscow group:
Network Infrared Bridge
6 PC compatible workstations
LAN server
LAN equipment (a rooter, switches, cables etc)
2*64 bin Opteron computation server
HDD Raid controller
Upgrades (HDD, controllers etc.)

The equipment purchased allowed to substantially enhance EIMB computational facilities.

Stipend
The following funds were allocated as stipends for the Russian participants

$5000 in 2004
$15,500 in 2006

The following scientists received the stipends from FIRCA grant:

Prof. V.G. Tumanyan; Drs V.J. Makeev, V.E.Ramensky, P.K. Vlasov, M.A. Roytberg, D.B. Malko, Y.V. Kravatsky; Technicians A.P. Lifanov, and A.I. Paskhin

Scientific outcome

1. **Compositional segmentation of the set of cytomegaloviri.** Compositional segmentation for viral genomes was done by the Basio program suit and was used by the Georgia Tech partners for more precise analysis of these genomes.
2. **Compositional segmentation of human genome.** Compositional segmentation of the sequences of all chromosomes of UCSC build 34 human genome was done by specifically developed scripts running Basio suit in the divide and conquer manner. Classification of the resulting segments into different compositional classes was performed for segmentation at different length scales and was used for training GeneMark.hmm models used for gene recognition in human genome
3. **Compositional segmentation of ENCODE gene set.** Compositional segmentation of gene segments selected for ENCODE project was done both for a sequence set with masked repeats and a set with repeats un-masked by the Basio suit. The segments were classified into compositional types. The resulting segmentation was used to build gene models used for gene prediction in the ENCODE project.
4. **Studying of periodical patterns in coding and non-coding regions of Drosophila genome.** Non-perfect tandem repeats have been identified in the protein-coding and non-coding regions of Drosophila. The ubiquitous repeating pattern with 6bp periodicity was identified. Strong triplet periodicity was identified in the coding regions.

Publications:

1. Valentina Boeva, Mireille Regnier, Dmitri Papatsenko and Vsevolod Makeev Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression, Bioinformatics 2006 22(6):676-684

Abstract:
Motivation: Genomic sequences are highly redundant and contain many types of repetitive DNA. Fuzzy tandem repeats (FTRs) are of particular interest. They are found in regulatory regions of eukaryotic genes and are reported to interact with transcription factors. However, accurate assessment of FTR occurrences in different genome segments requires specific algorithm for efficient FTR identification and classification.

Results: We have obtained formulas for P-values of FTR occurrence and developed an FTR identification algorithm implemented in TandemSWAN software. Using TandemSWAN we compared the structure and the occurrence of FTRs with short period length (up to 24 bp) in coding and non-coding regions including UTRs, heterochromatic, intergenic and enhancer sequences of Drosophila melanogaster and Drosophila

pseudoobscura. Tandems with period three and its multiples were found in coding segments, whereas FTRs with periods multiple of six are overrepresented in all non-coding segment. Periods equal to 5–7 and 11–14 were characteristic of the enhancer regions and other non-coding regions close to genes.

Availability: TandemSWAN web page, stand-alone version and documentation can be found at http://bioinform.genetika.ru/projects/swan/www/

2. V.J. Makeev, A.I. Paskhin, V.E.. Ramensky, D.B.. Malko, M.J. Borodovsky Compositional segmentation of *H. sapiens* genome and correlation of features with local composition. Proc. 2nd Int. Conf. MCCMB'05, July 18-21, 2005, Moscow, Russia

Abstract:
Genomes are known to be not homogeneous in their nucleotide composition at any length scale. Usually, the higher is the organization of the species, the more heterogeneous its genome. Many functional features correlate with the local nucleotide composition [1]. Despite many studies during the resent years, until known there have been no systematic classification of compositional structure of complete genomes of higher eukaryota with assessment of correlation of functional features with local nucleotide composition. Methods (i). Segmentation of sequences of complete genomes. We used our BASIO software [2] to perform segmentation of the sequences of complete genomes into domains with homogenous composition. BASIO includes one parameter, the boundary insertion penalty (BIP), which allows one to control the allowed heterogeneity of the resulting domains, the higher is the BIP value, the longer and more heterogeneous domain are obtained. In this study we interested mostly in the large scale segmentation, thus we used the lowest possible BIP value, with which the 3rd chromosome of Saccharomyces cerevisia did not exhibited internal boundaries after segmentation with BASIO. The 3rd yeast chromosome is known to be remarkably homogeneous [3] and can serve as a golden standard of a homogeneous domain on a genome length scale. The running time of BASIO software is proportional to the square of the input sequence length, which made impossible its direct application to the sequences of complete chromosomes. Thus we used divide-and-conquer strategy, with cutting the chromosome sequence into overlapping blocks and segmenting these blocks with an initial low BIP value with subsequent merging and re-segmentation of a whole sequence. A special attention was paid to a accurate consideration of non-sequenced regions of a genome, marked as long tracts of Ns in initial data. Finally, the segmentation of all chromosomes of UCSC build 34 Homo sapience genome was obtained. (ii) Segment classification. We have observed that often segments located at different chromosome regions or even at different chromosomes had a very close overall composition. Thus we performed clustering of segments into classes with a close nucleotide composition. To this end we used k-means clustering procedure implemented in MatLab software using as a distance measure the Eucledian distance in the compositional space. We tried classification into different number of classes, from 3 to 40. It was found, that for a low number of classes the vast majority of the genome sequence fell into a single class with other classes becoming outliers with a very abnormal composition (e.g. poly-A). On the other hand, the classification into a large number of classes yielded several large classes with a very close composition, which were almost not separated from each other in the compositional space. Finally, we adopted classification into 17 classes, from which 9 contained long segments from

different chromosomes and covered about 92.5% of the genome. The other 7.3% was covered by nonsequenced domains (poly-Ns). The remaining 7 classes covered in total only 0.05% of the genomes containing segments with abnormal composition. (iii) Feature mapping. We assessed such functional features of the genome as the total length covered with repeats, the average gene length, the total protein coding length, the average number of exons, the SNP density, the microsatellite density. It was found that clusters 8,9 were made almost entirely from simple repeats, and thus contained a small share of coding DNA. On the other hand a perfect anti-correlation between AT content and the overall coding DNA length was observed. SNP followed approximately the same pattern as the coding with some saturation trend, i.e. the 198 clusters with a percentage of coding sequences greater than 2% contained approximately the same SNP density, whereas for clusters with a less percentage of coding DNA, the SNP density increased linearly with the percentage of coding DNA. The additional consideration of A/T asymmetry demonstrates that 9 clusters form a remarkably symmetric pattern on the (A+T)/(A-T) plane with five clusters positioned along the line (AT~ 0) and four clusters positioned along the perpendicular line (A+T~0.63). It is also remarkable that three largest clusters containing 75% of total and 55% of coding DNA form a triangle near the crossing point A+T=0.63, A-T=0.

3 **T.V. Astakhova, S.V. Petrova, I.I. Tsitovich, M.A. Roytberg**
Recognition of coding regions in genome alignment
In: Bioinformatics of Genome Regulation and Structure II. (Eds. N.Kolchanov and R. Hofestaedt) Springer Science+Business Media, Inc. 2005, pp. 3-10

Abstract

Motivation:
Gene recognition is an old and important problem. Statistical and homology based methods work relatively well, if one try to find long exons or full genes, but are unable to recognize relatively short coding fragments. Genome alignments and study of synonymous and non-synonymous substitutions give a chance to overcome this drawback. Our aim is to propose a criterion to distinguish short coding and non-coding fragments of genome alignment and to create an algorithm to locate aligned coding regions.

Results: We have developed a method to locate aligned exons in a given alignment. First, we scan the alignment with a window of a fixed size (~40 bp) and assign to each window position P a score H(P). The value H(P) reflects, if numbers KS of synonymous substitutions, KN of non-synonymous substitutions and D of deleted symbols look like those for coding regions. Second, we mark "qualified exon-like" regions, QELRs, i.e. sequences of consecutive high-scoring windows. Presumably, each QELR contains one exon. Third, we point out an exon within every QELR. All the steps have to be performed twice, for the direct and invert complement chains independently. Finally, we compare predictions for two chains to exclude possible predictions of "exon shadows" on complementary chain instead of real exons. Tests have shown that ~ 93% of marked QELRs have intersections with real exons and ~ 93% of aligned annotated exons intersect marked QELRs. Total length of marked QELRs is ~ 1.30 of
total length of annotated exons. About 85% of total length of predicted exons belongs to annotated exons. The run-time of the algorithm is proportional to the length of a genome alignment.

Identification of Proteins Associated with Murine Cytomegalovirus Virions
Lisa M. Kattenhorn,[1] Ryan Mills,[2] Markus Wagner,[1] Alexandre Lomsadze,[3] Vsevolod Makeev,[4] Mark Borodovsky,[2,3]- Hidde L. Ploegh,[1] and Benedikt M. Kessler[1] Journal of Virology, October 2004, p. 11187-11197, Vol. 78, No. 20

Proteins associated with the murine cytomegalovirus (MCMV) viral particle were identified by a combined approach of proteomic and genomic methods. Purified MCMV virions were dissociated by complete denaturation and subjected to either separation by sodium dodecyl sulfate-polyacrylamide gel electrophoresis and in-gel digestion or treated directly by in-solution tryptic digestion. Peptides were separated by nanoflow liquid chromatography and analyzed by tandem mass spectrometry (LC-MS/MS). The MS/MS spectra obtained were searched against a database of MCMV open reading frames (ORFs) predicted to be protein coding by an MCMV-specific version of the gene prediction algorithm GeneMarkS. We identified 38 proteins from the capsid, tegument, glycoprotein, replication, and immunomodulatory protein families, as well as 20 genes of unknown function. Observed irregularities in coding potential suggested possible sequence errors in the 3'-proximal ends of m20 and M31. These errors were experimentally confirmed by sequencing analysis. The MS data further indicated the presence of peptides derived from the unannotated ORFs $ORF_{c225441-226898}$ (m166.5) and $ORF_{105932-106072}$. Immunoblot experiments confirmed expression of m166.5 during viral infection.

Ongoing projects

Improving quality of coding sequence prediction using genomic multiple alignments.
Identification of gene exon-intron structure, especially that of starting exons remains a major challenge in gene finding. One of the ways of increasing quality of gene recognition is usage of comparative genomics. Translated sequences of protein coding regions usually align better than those of non-coding regions, which can yield additional information on their positioning and boundaries. The objective of this project was to increase quality of gene finding in eukaryotes using genomic alignments.

Model organisms. For model organisms we selected 12 Drosophila genomes sequenced recently and available at http://rana.lbl.gov/drosophila/ and at UCSC genome website http://genome.ucsc.edu/.

Gene finding with GeneMark.hmm. Programs of the GeneMark family have an important feature that the HMM models for these programs can be derived from an anonymous genome by using automated self-training procedure. The set of Drosophila genomes is a convenient target for analysis. The genome of D. melanogaster, one of the best studied genomes currently sequenced, can be used for supervised training of gene recognition HMM. Genes can be predicted in other Drosophila genomes using GeneMark.hmm-ES self-training protocol that creates specific gene models. Genes predicted in D.melanogaster by both unsupervised and supervised version of the program can be compared with the high-quality annotation of this genome, which gives a reference point for prediction error for GeneMark.hmm. It should be noted that currently available genome annotations was built partly by homology to D. melanogaster genes, partly predicted in the genomes of Drosophila sp. using GenScan with gene models obtained for D. melanogaster. The models obtained via GeneMark self-training should in

principle perform better than those obtained with non-modified GenScan models. The secondary objective of the current project is to assess the improvement of annotation quality obtained via GeneMark model self-training. However, here the problem is that there are a limited number of experimentally verified genes in species other than D. melanogaster and thus we must limit ourselves with indirect evidences. One of such indirect evidences can be the quality of multiple alignments of predicted protein sequences.

The current stage of the project. In the current stage GeneMark.hmm-ES predictions were obtained for all 12 Drosophila sp. genomes. GeneMark.hmm prediction for D.melanogaster was compared to the experimental annotation and information on GeneMark annotation errors in the case of Drosophila genomes was obtained. The most important observation was that GeneMark.hmm-ES tends to overpredict genes in D. melanogaster species, particularly the very short short genes.

Future plans

Building multiple alignments of regions with predicted genes.
In the next stage our research we plan to build ortholog rows and multiple alignments of predicted genes. We are going to use Blat and Mercator software running at Georgia Tech computer facilities. Combination of Blat and Mercator was used to obtain multiple whole-genome orthology maps stored in UCSC genome database. In our case we are going to obtain the similar map, with the only difference that GeneMark.hm-ES annotations will be used instead of those currently available, which were made by GenScan. We are going to compare two annotations exon by exon and collect the statistics of differences between two sets of gene multiple alignments.

Predicting highly conservative segments upstream of the orthologous regions
Our next step will be prediction of highly conservative segments upstream of orthologous rows obtained with Mercator. To this end we are going to use SeSiMCMC Gibbs sampling algorithm, which can build multiple local alignment with one gap allowed. For the SeSiMCMC input we will supply the 5'-terminal segment of the orthologous regions extended upstream for several hundreds of base pairs. We expect that two possible types of aligned conserved regions will surfice. First, there can be parts of intial exons, including non-coding exon segments, missed for some reasons by combination of Blat and Mercator. Second, some non-transcribed conservative regions related to gene regulation can be detected. In the next step we will analyze this set of multiple local alignments of 5' regions obtained with SeSiMCMC.

Analyzing multiple alignments and obtaining improved annotation of initial exons.
We are going to assess features providing indirect evidence that a multiple local alignment contains undetected protein coding regions, or untraslated exons, or of untranscribed DNA. Protein coding regions should start with ATG codon downstream of which the alignment structure will have specific pattern typical for coding regions. Particularly, nucleotide substitutions are frequent in the silent codon positions, the sequence should contain no stop codons before first possible splice site and the observed codon usage should be compatible with one knwon for Drosophila. Aligned UTRs may contain motifs, characteristic for Drosophila genes. Finally, we will analyze the multiple alignments for the presence of putative transcription factor binding sites, which also can be responsible for sequence conservation.

The expected result Finally the improved annotation of starting exons in 12 Drosophila genomes will be obtained together with annotation of probable UTR and conserved binding sites for transcription factors.